A Proposed Methodology for Subjective Evaluation of Video and Text Summarization

Begona Garcia-Zapirain[1], Cristian Castillo[1], Aritz Badiola[1], Sofia Zahia[1], Amaia Mendez[1(✉)], David Langlois[2], Denis Jouvet[2], Juan-Manuel Torres[3], Mikołaj Leszczuk[4], and Kamel Smaili[2]

[1] eVida Research Group, University of Deusto, Bilbao, Spain
amala.mendez@deusto.es
[2] Loria, University of Lorraine, Lorraine, France
[3] LIA, Universite d'Avignon et des Pays de Vaucluse, Avignon, France
[4] AGH University of Science and Technology Kraków, Kraków, Poland

Abstract. To evaluate a system that automatically summarizes video files (image and audio), it should be taken into account how the system works and which are the part of the process that should be evaluated, as two main topics to be evaluated can be differentiated: the video summary and the text summary. So, in the present article it is presented a complete way in order to evaluate this type of systems efficiently. With this objective, the authors have performed two types of evaluation: objective and subjective (the main focus of this paper). The objective evaluation is mainly done automatically, using established and proven metrics or frameworks, but it may need in some way the participation of humans, while the subjective evaluation is based directly on the opinion of people, who evaluate the system by answering a set of questions, which are then processed in order to obtain the targeted conclusions. The obtained general results from both evaluation systems will provide valuable information about the completeness and coherence, as well as the correctness of the generated summarizations from different points of view, as the lexical, semantical, etc. perspective. Apart from providing information about the state of the art, it will be presented an experimental proposal too, including the parameters of the experiment and the evaluation methods to be applied.

Keywords: Video summarization · Objective and subjective evaluation
Text summary

1 Introduction and Literature Review

1.1 Introduction

AMIS is an original project concerning the second call: Human Language Understanding; Grounding Language Learning. This project acts on different data: video, audio and text. We consider the understanding process, to be the aptitude to capture the most important ideas contained in a media expressed in a foreign language, which would be compared to an equivalent document in the mother tongue of a user. In other words, the understanding will be approached by the global meaning of the content of a

K. Choroś et al. (Eds.): MISSI 2018, AISC 833, pp. 396–404, 2019.
https://doi.org/10.1007/978-3-319-98678-4_40

support and not by the meaning of each fragment of a video, audio or text. The idea of AMIS is to facilitate the comprehension of the huge amount of information available in TV shows, internet etc. One of the possibilities to reach this objective is to summarize the amount of information and then to translate it into the end-user language. Another objective of this project is to access to the underlying emotion or opinion contained in two medias. To do this, we propose to compare the opinion of two media supports, concerning the same topic, expressed in two different languages. The idea is to study the divergence and the convergence of opinions of two documents whatever their supports. Several skills are necessary to achieve this objective: video summarization, automatic speech recognition, machine translation, language modelling, sentiment-analysis, etc. Each of them, in our consortium, is treated by machine learning techniques; nevertheless human language processing is necessary for identifying the relevant opinions and for evaluating the quality of video, audio and text summarization by the end-user.

After analysing the existing different ways of evaluating an automatic summarizer system, and taking into account the objective of the actual evaluation, it is considered the best evaluation system a combination between subjective and objective evaluation methods. It is true that there are some automatic evaluation metrics/frameworks that have demonstrated good results, like ROUGE, QARLA… But these metrics/frameworks need human-generated summaries in order to compare, so, looking for the completeness of the evaluation, we consider necessary the inclusion of both perspectives of the analysis.

Besides, it should be considered that it is complicated to analyse automatic summaries with automatic systems using the lexical and phrase-based comparison with human generated summaries, obtaining really good proved results. Indeed, the humans, obviously, are not like machines; they have the ability to really understand, take the essence of something, and express it in a different way, in different words. So, it is possible to obtain a summary generated by a machine, and a summary generated by a human, and being in essence the same, but which can be expressed in a different way. In this case an automatic system probably will not be able to see the similarity in the words, the real meaning.

So, taking into account all the expressed ideas, the subjective evaluation is considered the best way to evaluate an automatic summarizer, and, to complete the evaluation from an objective perspective, it is interesting to apply some methods/metrics to analyze the obtained summaries. Anyway, both perspectives and their methods will be presented in the present document.

1.2 Literature Review

The state of the art of the evaluation of video summaries will be presented in the following tables from 2 perspectives: image and text. So, the most relevant and interesting papers about these topics are presented, including a resume about the most important part in relation with what we are analyzing, and some information about the parameters that are proposed (Table 1).

Table 1. Video summary evaluation.

Paper	Abstract	Experiment
Video Abstraction: A Systematic Review and Classification [1]	**Subjective:** user studies most useful and realistic	None
*VSUMM: A mechanism designed to produce static video summaries and a novel evaluation method [2]	**Objective with user participation:** In this evaluation method, called Comparison of User Summaries (***CUS***), the video summary is built manually by a number of users from the sampled frames. The user summaries are taken as a reference to be compared with the summaries obtained by different methods. In this way, the user summaries are the reference summaries, i.e., the ground-truth. Such comparisons are based on specific metrics, which are introduced in the following paragraphs	None
*A New Method for Static Video Summarization Using Local Descriptors and Video Temporal Segmentation [3]	**Objective:** ***CUS*** makes a comparison between the user summary and the automatic summary. The idea is to take a keyframe from the user summary and a keyframe from the automatic video summary	None
Automatic Evaluation Method For Rushes Summary Content [4]	**Subjective:** Each submitted summary was judged by three different human judges (assessors). An assessor was given the summary and a corresponding list of up 12 topics from the ground truth	- Users: 3 - 12 topics selected from the full video by the specialist

(*continued*)

Table 1. (*continued*)

Paper	Abstract	Experiment
Video Summarisation: A Conceptual Framework and Survey of the state of the art [5]	**Objective** methods do not incorporate user judgment into the evaluation criteria but evaluate the performance of a given technique based on, for example, the extent to which specific objects and events are accurately identified in the video stream and included in the video summary	- Users: 17 The users give feedback about the content, in terms of enjoyability and informativeness by means of informal discussions.
*A Pertinent Evaluation of Automatic Video Summary [6]	**Objective** (similar to ***CUS***): They propose an effective method for identifying the true matches between AT (Automatic Summary) and GT (Ground Truth User Summary) for the performance evaluation of the summarised videos. It includes the initial establishment of matched frames via two-way search followed by a consistency check where weak and false matches are eliminated	None
*Multi-video Summarization Based On Video-MMR [7]	- **Objective** (similar to ***CUS***): Is meaningful to compare Video-MMR (Maximal Marginal Relevance) to human choice? In a video set, 6 videos with most obvious features were chosen. Inside 6 videos, 3 videos own the largest distances with the others in this video set, while the other 3 videos have the smallest distances	- Users: 12 - Full videos: 3 Each user selects 10 keyframes from each video

(*continued*)

Table 1. (*continued*)

Paper	Abstract	Experiment
*VERT: Automatic Evaluation of Video Summaries [8]	By borrowing ideas from ROUGE and BLEU, the authors of this paper extend these measures to the domain of video summarization. We focus our approach on the selection of relevant keyframes, as a video skim can be easily constructed by concatenating video clips extracted around the selected keyframes… The authors talk about VERT-Precision and VERT-Recall, and how they are carried out	None
VSCAN: An Enhanced Video Summarization using Density-based Spatial Clustering [9]	In this paper, a **modified version** of an evaluation method Comparison of User Summaries (***CUS***) is used to evaluate the quality of video summaries. The modifications proposed to CUS method aims at providing a more perceptual assessment of the quality of the automatic video summaries	CUS (but more complete i our opinion)

2 Experiment Design - Participants and Protocol for the Whole Integrated System

Our proposed experiment includes two different lines: subjective evaluation (questionnaires) and objective evaluation.

In order to do the evaluation as complete as possible, and obtain the best results, the problem should be approached from different perspectives, so, it has been combined information from different sources [13, 14] and developed a new way of evaluation in order to obtain better results. On each perspective there are some questions with a specific format for the answer, which can be multiple choice or ranking from 0 to 4 formats. In the multiple choice format, some specific answers will be provided in order to be selected one or some of them by the user. In the case of the scoring from 0 to 4,

Table 2. Questionnaire for the video and text summary evaluation

Criteria	Excellent	Very good	Good	Fair	Not done	Comments/Suggestions
Summary video						
Is the summary understandable?	4	3	2	1	0	
The video doesn't contain any part out of context, or it does not affect to the main expressed ideas	4	3	2	1	0	
***Different questions* about the original video, in order to ensure that the summary contains the key ideas and the user is able to get these ideas from the summary**	See below one example.					
Summary text						
Is the summary understandable?	4	3	2	1	0	
Is it lexically/grammatically correct?	4	3	2	1	0	
Is it semantically correct?	4	3	2	1	0	
Does it contain redundant information?	4	3	2	1	0	
Are the references (it, she, he...) clear? (*looking for lack of information*)	4	3	2	1	0	
***Different questions* about the original video, in order to ensure that the summary contains the key ideas and the user is able to get these ideas from the summary**						
Summary video and summary text						
Do you think that both, the summary video and the summary text, express the same idea? (*cohesion between the both formats is measured*)	4	3	2	1	0	

each number has a meaning: 0 = not done, 1 = fair, 2 = good, 3 = very good and 4 = excellent. The subjective evaluation can be done in 2 ways:

(a) Assessment of Summarized Video (image sequence) and Text

Participants: 25 per language (Arabic and French) (balanced number of men and women)
Inclusion criteria: men and women over 18 years old, with at least high school level.
Exclusion criteria: reading and writing impairment. Understanding problems.
Number of summarized video and text per user: Every user will review **a** set of 3 videos (out of the 25 prepared) with mixed topics. The test will be made of the summarized video and text version (in English) (Table 2).

(b) Assessment of the Coherence between Original and Summarized Video and Text

Participants: Four in Arabic and four in French (two men and two women for each language)
Inclusion criteria: men and women over 18 years old, with at least high school level. The participants have to be fluent in both languages.
Exclusion criteria: reading and writing impairment. Understanding problems.
Number of summarized video and text per user: one with mixed topics. One video will be selected per language for the test with original (Arabic or French) and summarized video and text version (English) (Table 3).

Table 3. Questionnaire for original and summarized video and text evaluation

Criteria	Excellent	Very good	Good	Fair	Not done	Comments/Suggestions
Original and summarized video and text						
Do you think that the provided summarizations, in video and in text, really express the main ideas of the original video?	4	3	2	1	0	
Can they be considered good summaries? **- Video** **- Text**	4	3	2	1	0	
Are they long enough to contain the main ideas? **- Video** **- Text**	4	3	2	1	0	
Do you consider it too long?	4	3	2	1	0	

Below, we propose some examples of evaluation with original videos. Specific questions are asked for each video:

The-rise-of-chemsex-on-Londons-gay-scene---**BBC-News**	**Criteria**
Which one is the main topic of the summary?	1) Homosexual parties 2) Homosexuality and STDs 3) Homosexuality, drugs and STDs
The summary says that there is not any relation between drugs and STD:	1) True 2) False
The diagnosis of HIV with drugs is related?	1) True 2) False

The assessment data analysis will consist on statistical analysis of questionnaires and the application of some machine learning techniques if possible for clusterization and comparison purposes between genders, language.

3 Results

The results regarding the questionnaires would be based in a point system, and the criteria of quality, or the different point ranges of quality level will be established depending on the total number of questions.

During 2018, all the proposed test will be carried out.

Acknowledgements. Research work funded by the Spanish Ministry of Economy, Competitiveness and Industry (Spain) conferred under the Chist-Era AMIS project.

References

1. Truong, B.T., Venkatesh, S.: Video Abstraction: A Systematic Review and Classification (2007)
2. Fontes de Avila, S.E., Brandão Lopes, A., da Luz Jr, A., de Alburquerque Araújo, A.: VSUMM: a mechanism designed to produce static video summaries and a novel evaluation method (2010)
3. Cayllahua Cahuina, E.J., Camara Chavez, G.: A New Method for Static Video Summarization Using Local Descriptors and Video Temporal Segmentation (2013)
4. Dumont, E., Bernard, M.: Automatic Evaluation Method for Rushes Summary Content (2009)
5. Money, A.G., Agius, H.: Video Summarisation: A Conceptual Framework and Survey of the State of the Art (2008)

6. Kannappan, S., Liu, Y., Tiddeman, B.: A Pertinent Evaluation of Automatic Video Summary (2016)
7. Li, Y., Merialdo, B.: Multi-video Summarization Based on Video-MMR (2010)
8. Li, Y., Merialdo, B.: VERT: Automatic Evaluation of Video Summaries (2010)
9. Mohamed, K.M., Ismail, M.A., Ghanem, N.M.: VSCAN: An Enhanced Video Summarization using Density-based Spatial Clustering (2014)
10. Molina, A., Torres-Moreno, J.-M.: The Turing Test for Automatic Text Summarization Evaluation (2016)
11. Molina Villegas, A., Torres-Moreno, J.-M., Sanjuan, E.: A Turing Test to Evaluate a Complex Summarization Task (2013)
12. Lin, C.-Y., Hovy, E.: Manual and Automatic Evaluation of Summaries (2002)
13. Hassel, M.: Evaluation of Automatic Text Summarization: a practical implementation (2004)
14. Saziyabegum, S., Sajja, P.S.: Review on Text Summarization Evaluation Method (2017)